MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2017/2018

TDS2101 – INTRODUCTION TO DATA SCIENCE

(All sections / Groups)

28 OCTOBER 2017
9.00 a.m. – 11.00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 4 pages with 4 Questions only.
2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

Question 1

a) State ONE key responsibility of data scientists.

(1 mark)

b) Big data is commonly characterized by the 5V's. Describe value, velocity and veracity.

(3 marks)

a) Briefly explain how Target (a US-based retail company) and Facebook (social media) leveraged on Big Data.

(2 marks)

c) Differentiate between traditional data analysis implementations compared to modern data analysis (Big Data) in terms of (i) Data sources and (ii) Data structures.

(4 marks)

Question 2

a) A data scientist formulated the question "What is the relationship between a range of dietary factors and viral illnesses?" as part of his analysis.

i. Which *type of question* (among the six types of questions) has the Data Scientist formulated?

ii. If the Data Scientist wants to study the impact of changing one factor in the analysis, which type of question should he formulate?

(2 marks)

b) The dataset the time spent by students on three key activities i.e. study, sleep and sports. The values represent the average hours for study per day, average hours of sleep per day and percentage (%) of time spent on sports in a week.

		1	2	3
	Name	**Study (hours/day)**	**Sleep (hours/day)**	**Sports (%/week)**
1	Alisha	5	4	7
2	Melissa	12	5	7
3	Ivan	6	7	6
4	Ridzwean	4	6	6
5	Joe	25	8	10
6	Farid	16		10
7	Will	2	9	4
8	Zhen	27	6	2
9	Gan	2	12	5
10	Kelly	10		5

Continue...

i. Identify any outliers within the *study* attribute (indicate the row and column number (row:column)). Discuss whether the outliers are likely to be errors or anomalies.

ii. Identify any errors/inconsistencies in the *sleep* attribute (indicate the row and column number (row:column)). Determine which data quality dimension does this vilotate and discuss ONE approach that can be adopted to handle this error/inconsistency.

(5 marks)

c) Data quality dimensions must be adhered to when dealing with Big Data.

i. Describe the *validity* data quality dimension.

ii. Explain the issues that may arise when trying to ensure *validity* for the "name" attribute in Malaysia which is a multiracial community.

(3 marks)

Question 3

a) Write the Python code required to retrieve rows and columns from a *DataFrame* using the *iloc* method. You do not need to write a full program.

i. Retrieve the entire *DataFrame*

ii. Retrieve the first 5 rows, and all of the columns for those rows.

iii. Retrieve the first column, and all of the rows for the column.

iv. Retrieve the 10th row, and all of the columns for that row.

(4 marks)

b) The CEO of a hypermarket is considering rearranging the position of items on the sales racks in an attempt to boost sales of under-performing items. As the data scientist, how can you help the CEO decide which items should be placed close to each other. Explain how you would handle this analysis and state the choice of model/algorithm you may need to use for this purpose.

(2 marks)

c) Describe two possible attributes that can be correlated but do not have a causal relationship. Justify your answer.

(4 marks)

Question 4

a) Differentiate between SQL and No-SQL databases in terms of the (i) storage model and (ii) schema

(4 marks)

b) Suggest ONE reason why R may be a preferred language for Data Science projects.

(2 marks)

Continue...

c) A junior data scientist is seeking your advice on when to use graphs instead of tables to visualize results. Discuss what should be considered when making a decision on the type of representation (graph or table) to be used.

(4 marks)

End of Paper